COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

USENIX

SoK: PHILTER: Uncovering Security and Functional Gaps in AI-based Phishing Website Detection Literature via an LLM-based Reasoning Framework

Mahbub Alam[1], Muhammad Lutfor Rahman[2], Sonjoy Kumar Paul[1], Amy W. Hays[1], Aftab Hussain[1], Md Imanul Huq[1], and Nitesh Saxena[1]

[1]Texas A&M University, [2]California State University San Marcos

SPIES
Security and Privacy in Emerging computing and networking Systems

High detection accuracy did not solve phishing

- Phishing detection papers often report accuracy over 98%.
- Phishing is the most reported cybercrime (FBI IC3).
- It's a gateway to major cybercrimes like ransomware and data breaches.

191K

Phishing Incidents

Reported to FBI IC3 in 2025

$20.9B

Total Cybercrime Losses in 2025

Up from $16.6B in 2024

FBI IC3 Annual Reports –– www.ic3.gov/annualreport/reports

Beyond Accuracy

Despite high benchmark accuracy, deployment readiness remains unclear:

Functionality

Phishing Tactic Diversity: Which tactics does it catch, and which slip through?
Benign Page Diversity: Is it tested on the benign pages users actually visit?
Interpretability: Does it explain why a page was flagged?
Evaluation Transparency: Does it report metrics suited to class imbalance?

Security

Concept Drift: Does it adapt as attackers evolve?
Active Attacks: Can it withstand evasion attempts?
Privacy: Does it protect user data?

PHILTER evaluates 55 papers against these seven requirements, guided by a codebook.

Study Selection and Taxonomy

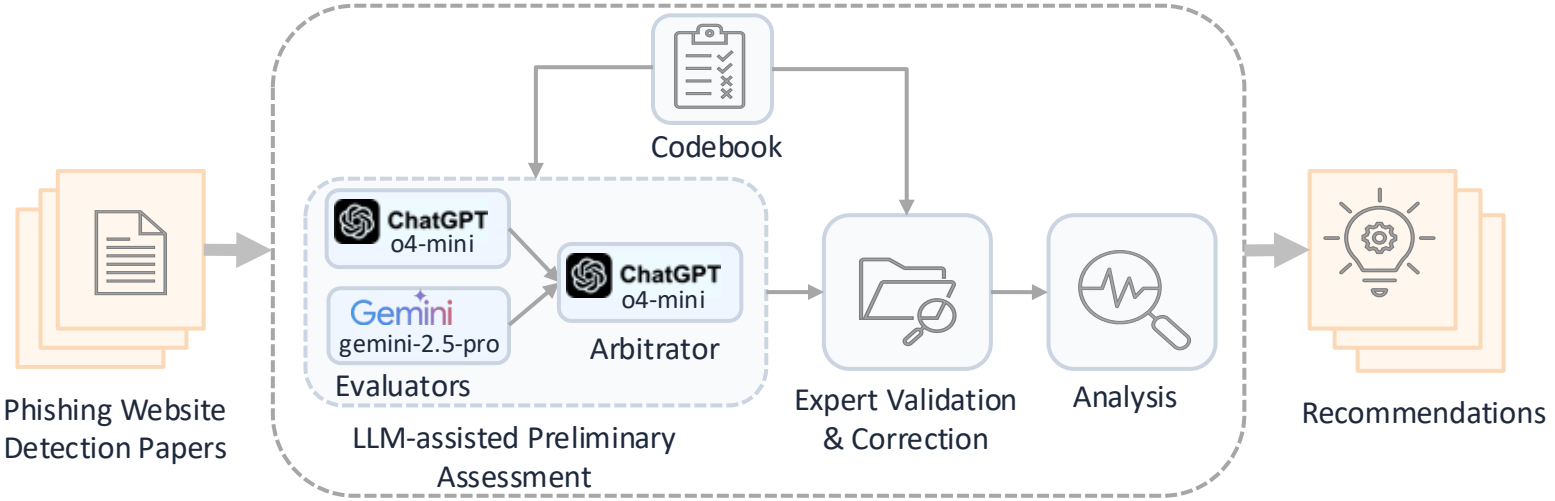
PRISMA-Guided Paper Selection

- **Keywords:** *phishing (url|page|web|site) (detect|attack|adversarial)*
- **Initial corpus:** 288 peer-reviewed articles from 178 venues (2019-2025).
- **Final corpus:** 55 papers after PRISMA-guided screening, venue/citation filtering, and snowballing.

Detection Strategy Taxonomy

- **Feature-based:** ML on extracted URL/page features.
- **Similarity-based:** Compares visual/textual similarity with reference sites.
- **Identity-based:** Infers claimed brand and verifies brand-domain consistency.
- **Hybrid:** Combines multiple detection strategies.

PHILTER Framework



- 89% LLM–expert agreement, independently validated by 2 authors.

Key Findings

Feature-based	Similarity-based	Identity-based	Hybrid
43% High	**43% High**	**100% High**	**44% High**
Privacy preservation	**Interpretability**	**Interpretability**	**Drift adaptation**
Lightweight client-side detection	Flags suspicious page elements	Brand-domain matching	Complementary strengths
83% Low	**86% Medium**	**100% Low**	**89% Low**
Benign page diversity	**Drift adaptation**	**Privacy preservation**	**Privacy preservation**
Popular domain bias	Static reference sets, temporal evaluation	External lookups	External lookups

Key Findings

Functionality Gaps

84%
only partial

Limited tactic diversity
No per-tactic evaluation — it remains unclear which tactics are detected, and which are missed.

75%
rated low

False positive gaps
Often exclude evaluation on low-reputation domains and legitimate pages with embedded brand elements.

53%
rated low

Underexplored interpretability
Most feature-based methods act as black boxes, limiting transparency and trust.

Security Gaps

49%
rated low

Limited drift adaptation
Assume phishing tactics stay static, with no retraining or adaptive mechanism.

55%
rated low

Limited attack resistance
Limited evaluation against evasive manipulations.

67%
rated low

Privacy gaps
Sends full URLs, page content, or screenshots to external servers.

No paper satisfies all functionality and security requirements

Recommendations

Functionality

Tactic-aware evaluation: Build tactic-annotated phishing benchmarks and show per-tactic performance.

Benign-page generalization: Test false positives on regional sites and pages with third-party logos.

Enhance interpretability: Surface suspicious page elements, visual cues, or brand-domain mismatches.

Security

Strengthen drift adaptation: Develop drift-adaptive mechanisms (auto refreshed reference list, incremental learning, drift-resilient features) and test with temporal split.

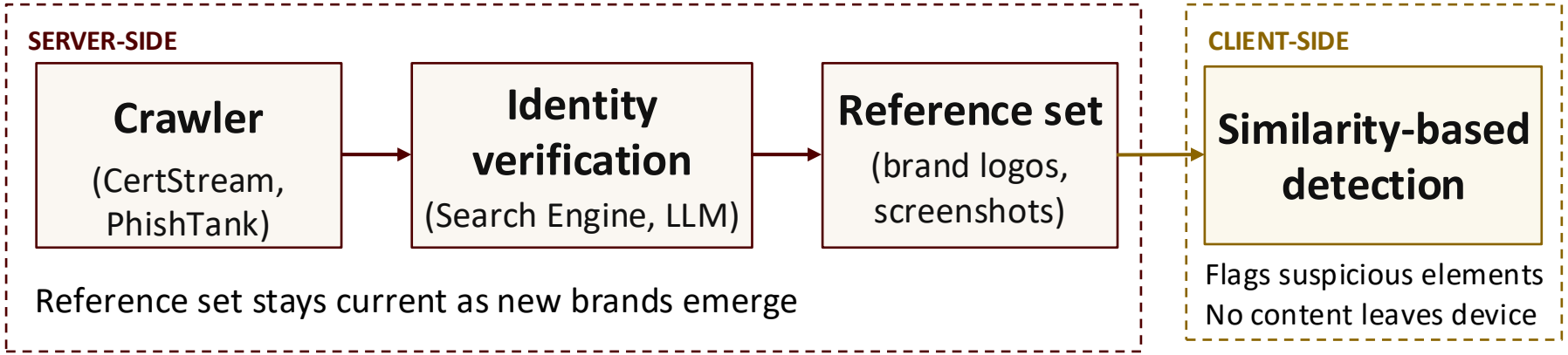
Strengthen attack resistance: Evaluate against evasive manipulations (obfuscated HTML, manipulated visuals, interaction-gated content) and adversarial attacks.

Privacy-preserving design: On-device detection, federated learning, and anonymized feature sharing.

Integrate complementary strengths from different strategies.

Recommendations

An illustrative design combining complementary strengths:



Drift-adaptive, privacy-preserving, and interpretable by design

Thanks for Listening!

Check out **PHILTER** code and artifacts.

Reach out: mahbub.alam@tamu.edu

SPIES
Security and Privacy In Emerging computing and networking Systems